Understanding Morality behind Human opinion and reasoning

Akit Kumar
Syracuse University
akumar32@syr.edu

Lu Xiao
Syracuse University
lxiao04@syr.edu

Abstract

Moral values influence people's opinions and reasoning. While people increasingly share and shape each other's opinions in social media, we have limited understanding of how people's moral values play a role in these processes. As a first step to close this literature gap, we conducted two studies to examine the relationships between one's morality and how one's opinions are expressed and changed in social media. Specifically, we explored the potential of using the moral values reflected from an online comment to classify its stance. We also examined how these moral values contribute to the change of one's opinions in online persuasion processes. Our results show that one's morality and viewpoint are connected, and the more similar between the two users' moral values the more persuasion power one carries over the other. In addition, the stronger a user's moral value the more resistant the user is to change their opinion.

1. Introduction

With the introduction of social networking websites and their continued development and growth, many people rely heavily on social media to express and reason their thoughts towards social issues and justice. Shared almost instantly and with a large number of people, these users generated content on social media can have significant influences on social, political, economic, and other potentially contentious matters and can lead to disagreements, abusive discussions, and separated communities in the society [1]. Social media environments have become a common and important place in the society for people to share and shape each other's opinions. Of many factors that influence people's opinions such as people's educational background and political views, their moral values and judgements are key influencing aspects. While these opinions are shared widely in social media, we have limited understanding of how people's moral values play a role in expressing and changing their opinions.

Interested in closing this literature gap, we conducted two studies to explore the above two aspects. Specifically, social media communications are often in the form of textual data and reflect their internal reasoning processes [2]. We applied computational text analysis techniques to examine the relationships among people's moral beliefs reflected from their online textual comments, their opinions expressed in the comments, and their acknowledgement of the opinion change. Using this methodology, the first task we encountered in our exploration is to detect morality in the user-generated content. Detecting morality in a text necessitates trustworthy operationalization and detection of regularities [3]. A couple of dictionaries are developed for this purpose: Moral foundation Dictionary (MFD) [4]: Enhanced Moral foundation Dictionary (Enh MFD 1) [5], and Enhanced Moral foundation Dictionary (Enh MFD 2) [6], and eMFDScore tool [7]. Leveraging these resources, we used these morality lexicons in detecting morality in the texts in this study.

In the first study, we explored the connection between one's moral beliefs and the expression of their opinions in social media. We asked this question: does the morality of the user, reflected from the user's comment, contributes to the classification of the user's stance regarding an issue? Automatically detecting whether the author of a piece of text is in favor of or against a certain objective has been termed as stance detection. In extant research [6], support vector machine (SVM), random forest (RF) classic models, and Long short term memory (LSTM) deep learning models have been trained on Semeval 2016 Stance detection benchmark dataset [8]. With this dataset and the machine learning approach, we experimented various Morality Feature selection techniques using MFD, Enh MFD 1, and Enh MFD 2. Our model that includes the morality features achieved a highly competitive performance in stance classification, compared to the best performing stance classification model reported in

URI: https://hdl.handle.net/10125/79708
978-0-9981331-5-7

the literature [9, 10]. In [9]'s study, the researchers used the Semeval 2016 dataset in their algorithm and achieved an F1 score of 0.781 for tweet based trained BERT model. Trained on SemEval 16 dataset, our algorithm achieved a F1 score of 0.89 when trained using morality focus words. [10] used the same dataset in their training and achieved an F-score of 0.7.

In the second study, we examined that when faced with the persuasion attempts in an online discussion, if and how the user's morality, reflected from their argument, affects the process of changing their opinions. While prior experiments [11] and persuasion suggest the connection between one's morality and their responses to persuasion attempts [12], there is an insufficient investigation about this connection in online communication contexts. Our second study is an attempt to close this research gap. The online discussions we studied are subreddit changemyview (CMV) discussions. These discussions are often examined in online persuasion research [11]. In this subreddit, the original post from a user is about the user's argument regarding an issue. The others offer their viewpoints with the intention to change this user's perspective. Situated in the CMV discussion context, we analyzed the morality similarity between the original post and the comment that changed user's view (persuasive comment). The results show that a persuasive comment is more similar to the original post than a non-persuasive comment in the morality aspect. This finding is consistent with earlier studies that arguments toward more liberal or conservative moral values tend to be more persuasive to liberals or conservatives respectively [13].

Prior studies also suggest the connection between morality and resistance to persuasion. In general, stronger morally based attitudes are particularly resistant to persuasion and can result in the rejection of disagreeing with others [12]. From this aspect, we examined whether and how the strength of the original poster's morality correlates with when they tend to issue delta points during a discussion. Our analysis shows that regardless of whether the moral dimension is negative or positive, people with higher moral value in their original post tend not to issue delta points until later in the discussion.

2. Related Work

2.1. Definition of Morality and Measurement

Morality in general is known as our feeling towards what is right and wrong. It could be our instinct to protect others (care), the ideology of justice, right, and autonomy (justice), our instinct of punishment for incorrect (Fairness), or it could be our virtues of patriotism and self-sacrifice for the group (loyalty). The most recent advent of pragmatic utility, Moral Foundations Theory (MFT) [4] states that morality highly varies across cultures yet shows many similarities. These similarities are society's core values and can be divided into five core categories. In each category, there are two directions: the virtue direction that represents the moral excellence in the category and the vice that shows the depravity of the moral category. Table 1 presents these five categories and the directions of virtue/vice.

Moral Foundations Theory has been applied in studies that attempt to identify one's moral values based on one's writing. Specifically, Moral Foundation Dictionary (MFD) was created based on the theory [14]. It contains 32 words that are used in day-to-day life for each category. MFD has been used in various studies [6, 15]. Recently, this dictionary has been enhanced by different research groups. Frimer et al. increased the count from 32 words to 210 words per category of MFD (Enh MFD 2) using word2vec software [5] Rezapour and colleagues proposed an enhanced MFD increasing the original MFD count from 324 to 4,636 [6] through a semi-automated and human-validated approach. This created an enhanced MFD (Enh MFD 1).

In the previous approaches trained experts identified morality-related content independent of the context, Differently, Hopp [7] developed the extended Moral foundation Dictionary (eMFDScore) based on a crowd-sourced text highlighting task [3]. Used for extracting morally relevant information from real-world messages, eMFDScore combines natural language processing techniques with basic, spontaneous responses to moral content in text. eMFDscore provides metrics for analyzing moral information, and extracts moral patient, agent, and attribute words related to entities.

2.2. Morality and expressing Opinions in social media

People express their opinions on social media. Opinion mining, also known as sentiment analysis, is a text analysis methodology that employs computational linguistics and natural language processing to detect and extract opinion from text automatically (positive, negative, neutral, etc.). Models for opinion mining and sentiment analysis can focus on the polarity of opinion (positive, negative, or neutral), personal feelings (angry, glad, sad, etc.), and intentions or goals (interested or not interested). In this research, we are focusing on

Table 1. Moral Foundation Theory

Category	Virtue / Vice
Ability to understand the pain of others	Care / Harm
Conceptualization of justice, right, and autonomy	Fairness/Cheating
Human socialization concepts like patriotism, and self-sacrifice towards society	Loyalty/Betrayal
Endorsing hierarchical social structure	Authority/ Subversion
Promotes the psychology of disgust and contamination, and cleanliness of the body	Sanctity/ Degradation

stance detection with the 3-class problem (Against, Favor, and None). Initial work on stance detection focused on preliminary debates [16] where transcripts of U.S. congressional floor debates are used to assess whether speeches signify support for or opposition to proposed legislation. This paved the way for using stance detection as the core component of fact-checking [17], fake news detection [18], and rumor verification [19]. However, recent research has primarily focused on single datasets and domains like SemEval 2016 dataset [8] to train various models [6, 8, 10] to get the best prediction results. Recent research done by Benjamin, S. and team [10], shows prediction accuracy of stance detection on SemEval dataset using various models on ten different datasets.

Basic principles of human values and the expression of opinion in a text are related [8], Hence considering morality as a basic principle of human value helps us better understand the opinions in the user-generated text [6]. The majority of earlier studies focus on Morality used in analyzing social effects and causes using MFD. Studies have explored MFD in socio-political disputes[20], latent semantic analysis to assess morality in tweets concerning a variety of social topics such as homosexuality and immigration [21], and assess polarized conversations in news sources [22].

A study done by Rezvaneh, R. [6] worked on the assumption that People's values are reflected in their language use and are influenced by their cultural contexts [23]. They used the Enh MFD 1 and captured the digital traces of human behavior by measuring abstract and complex constructs such as personal values and social effects (Morality and Stance). Their findings show that in the vast majority of circumstances, the Moral Foundations Dictionary improves prediction accuracy, particularly when utilized for feature-based machine learning. The moral value of a sentence is based on its complete context. The limitation of Traditional feature-based machine learning models, and deep learning models (RNN with LSTM, RF, and, SVM), used in the existing work, are sequence-based models and fail to account for sentence context when considering moral values, and hence cannot fully grasp sentences' morality features. On the other hand, Transformer-based models such as Bidirectional Encoder Representation from Transformers (BERT) [24] and XLNet [25] use an attention mechanism that learns contextual relations between words in a sentence. Therefore, in this study, we have used various MFD's to extract moral features and used them to train Transformer-based models.

2.3. Morality and Changing Opinions in social media

More abstract moral concepts that can be utilized to regulate and govern the interactions of individuals in larger and more complicated societies have been addressed by philosophers, legal experts, and political scientists [4]. The essence of cooperative or empathic conduct is far more symbolic, as it is based on more abstract and ambiguous ideas such as "the greater good," rather than on direct exchanges between specific persons. These ideologies help society interact more effectively and grow in unity and harmony. However, research shows that achieving social welfare and success is dependent on one's ability to persuade and influence others [26]. Persuasion is "human communication that is designed to influence others by modifying their beliefs, values, or attitudes" [26] and there is a relationship between human belief, values, or attitudes and human morality [6]. Following this line of research, we have assumed that humans use similar moral content to become better persuaders in a society. To validate the hypothesis, we have used the Sampling distribution and Hypothesis test using P-test.

This approach gives us insight into the morality aspect of persuasion. On the other hand, recent evidence in the area of morality shows that morally based attitudes are particularly resistant to persuasion and can result in the rejection of disagreeing others [27] and people with stronger moral attitudes are more resistant to change [12]. Hence, we can assume that if OP has a high moral attitude towards the vice/ virtue of MFD, the OP might resist the change of its opinion in the initial Comments. Following the earlier studies, we have come up with two questions, does Morality similarity play a role in persuasion on social networks, and are people with stronger moral attitudes more resistant to change?

3. Research Methodology

3.1. Measurement of Morality Dimensions

We extract morality from the input text using three different Feature selection techniques [6]:

- Morality Type: For each input text we match words in the vice or virtue category of each morality dimension. This forms a vector of size 10 and each dimension of the vector represents each category of MFT.
- Morality Dimension: For each input text we match words in any of the five morality dimensions. A vector of size 5 is generated using this approach where each dimension of the vector represents one moral dimension.
- Morality Polarity: The input text has weights on vice and virtue category of MFT. This approach presents the text in a vector of size 2

Apart from the number of moral words per input text in the above approach, we have used the eMFDScore tool [7] to capture the moral context in an input text. In the eMFDScore lexicon containing 3020 words, each word has 5 probabilities that denote the likelihood of the word being associated with the five moral foundations as identified in MFT. For example, the word "kill" in the lexicon has care probability as 0.4 and loyalty probability as 0.24 which means a 40% chance that a coder highlighted that a context containing the "kill" word as the care-harm foundation and 24% chance as the loyalty-betrayal foundation. Apart from 5 categories, each word has been assigned a 5 sentimental score that denotes the average sentiment towards each moral category of the foundational context in which the word appeared. Based on these scores, the eMFDScore tool generates a vector of size 10 for each input data.

3.2. Research context

The first study explores the connection between one's morality and their expression of opinion. We experimented with different morality lexicons including MFD, EnhMFD-1, EnhMFD-2 to extract the morality features from textual data and used Feature selection techniques to generate moral vectors. SVM, Random Forest, and RNN with LSTM, and Transformer based models were trained using moral vectors to emphasize on morality features of the data. While our second study examines two aspects of the relationship between morality and change in opinion (persuasion). We examine whether and how the similarity between persuader and persuadee in terms of their moral beliefs, that is, their morality similarity, plays a role in the persuasion process. In addition, we explore this connection in the online persuasion processes through the analysis of people's online arguments in response to the persuasion attempts. Figure 1 Is a flow diagram which depicts the phases involved in the two studies.

Table 2. Descriptive statistics of SemEval 2016 stance dataset

Number of Tweets	Abortion	Atheism	Climate	Clinton	Trump	Feminist
Against	544	464	26	565	299	511
Favor	167	124	335	163	148	268
None	222	145	203	256	260	170
Total	933	733	564	984	707	949

3.3. Datasets

For the first study, we have used SemEval 2016 dataset [8] as earlier studies have used SemEval 2016 dataset for training various machine learning models for stance prediction [6, 10]. This dataset contains 4,870 tweets on six issues: abortion, atheism, climate change, feminism, Donald Trump, and Hillary Clinton. The tweets were hand-coded for a political viewpoint, with options such as in favor, against, and none. Table 2 displays the number of tweets as per 6 issues for Against, Favor, and None stance.

For the second study, the dataset was collected from the CMV forum from January 2014 to May 2015 [28]. CMV is an active subreddit forum with 1.3m active users, where the user provides their views on a subject either in the comments or as an Original Poster (OP). The rule of the forum is that OP creates a post about a topic in which OP needs to explain the topic and the reasoning behind its view on the topic. The post is required to at least have 500 characters. Once a post is created, the other users of the forum can present their views on the topic to try to persuade the initial views of the OP. CMV has rigid moderation regarding the comments provided by the users to OP, like the direct comment to OP post "must challenge at least one aspect of OP's current view (however minor)". It also forbids the use of obscene or abusive language and condemning or blaming others for their unwillingness to modify their minds. When a comment successfully alters an OP's perspective, CMV expects the OP to give a delta point (DP) and explain why and how the comment affected his or her perspective. These guidelines and moderation make the CMV forum a unique and valuable place for studying online persuasion. The CMV dataset in our

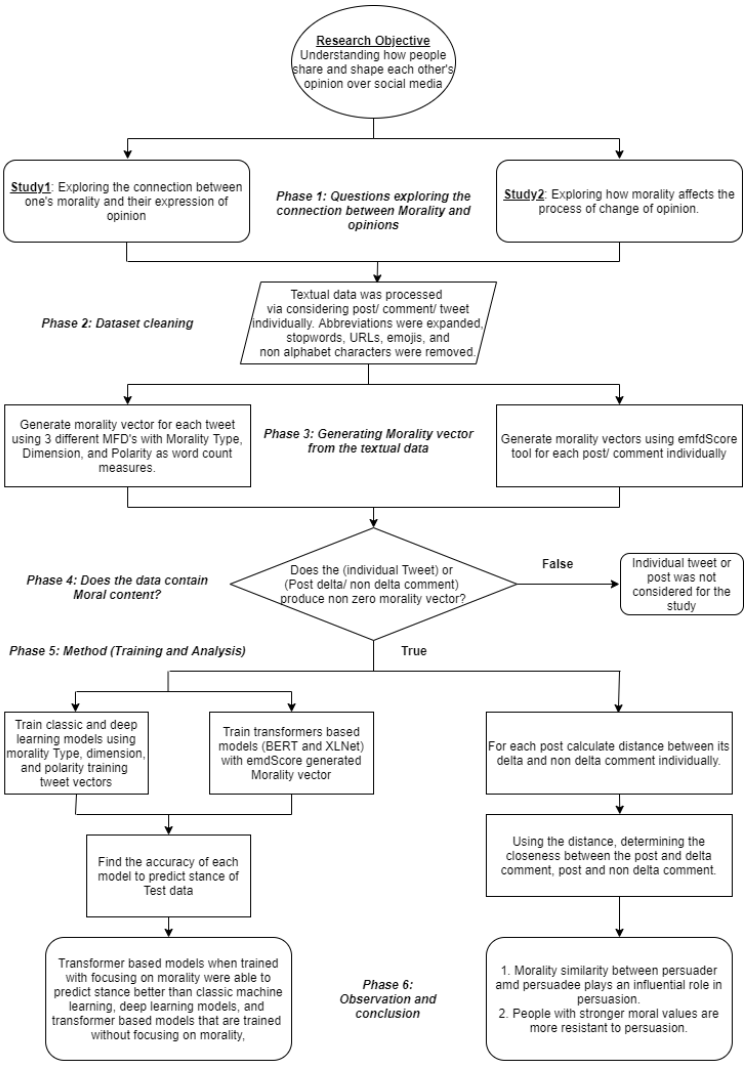

Figure 1. Flow diagram describing the phases involved in the two studies

study is divided into two files: The original post and the Comments. Original post dataset contains 4600 Original posts and Comments dataset 110251 comments. Each original post has 1 or more comments with delta points and is considered as persuasive comments for the OP's to change their mindset. Comments which don't have delta point are considered as non-persuasive for this research. CMV comments can occur in the form of threaded discussions, in which users can respond to each other's remarks and create a conversation tree. At different levels of the tree, a delta can be given to comments. In this research, we are focusing on the morality of the persuasive comment and hence we are considering all the comments despite their comment Level.

The data cleaning process for both the datasets are similar. The preprocessing of the data includes removing the noisy and unstructured textual data. This is done by expanding all the abbreviations and then removing all punctuations, URLs, hashtag symbols, usernames, numbers, and special characters Finally the textual data was converted to lowercase.

After the data cleaning process, in our first study, we have used the Bag of word approach with feature selection techniques to create a morality vector that represents the morality of the input text. The moral words were taken from MFD, Enh MFD 1, and Enh MFD2 individually. Morality vectors were used to train machine learning, deep learning models, and transformer-based models. In our second study, we treated each post individually and calculated the morality vector using the eMFDScore tool. The output morality vector for each post will be of size 10, 5 moral categories, and 5 sentimental categories. We used the

same approach to calculate the morality of comments. The input data that generates a morality vector with all zero values, is not considered for our studies.

3.4. Method (Training and Analysis)

3.4.1. Study 1: Morality and Expressing Opinions

Taking the input's morality aspect into consideration, Rezvaneh R. and team leveraged MFD in their supervised classifiers for stance classification [6]. Various machine learning approaches were experimented in their work, including SVM, Random Forest, and RNN with LSTM. In our study, we experimented with different morality lexicons including MFD, EnhMFD-1, EnhMFD-2. For a better comparison with their study, we applied the same machine learning algorithms and adopted the same parameters reported in their paper. For instance, we separated the dataset into its original sub-topics and trained the SVM and RF models based on individual sub-topics using the Python Scikitlearn package [29]. With the SemEval 2016 stance dataset, the split between training and testing data is also 80:20. We also investigated the deep learning model RNN with LSTM. The embedding layer of LSTM is created with an intersection between the words from the lexicons and the dataset. Morality type, morality dimension, and morality polarity are not considered here. This embedding layer was then created using the 200-dimensional word embedding from GloVe Twitter trained on two billion tweets [30]. A Bidirectional LSTM of sizes 100 and 150 were formed with a hidden layer with a Sigmoid activation function, and an output layer with Softmax activation function followed the embedding layer. The parameters were further optimized using Adam [31] with cross-entropy as the loss Function.

We have used two transformer-based models BERT (bert-base-uncased) and XLNet (xlnet-base-cased) and trained them based on the Semeval 2016 dataset with 80:20 training and testing ratio while using Multiple MFD's to focus onmoral words. BERT and XLNet models have been pre-trained using a huge corpus of data with predefined weights in each layer. When training the BERT model with SemEval data, we found that the number of words commons between morality lexicon and Bert tokenize vocab is 1242 out of 4419 and the dataset was too small to train the model while focusing on moral words. Therefore, when testing the model's accuracy, the model will take weights of uncommon words from previously trained data. To overcome this, we used those sentences which had at least one word from the morality lexicon. Each sentence having a morality word was trained along with its left and right sentence for the complete context of the training sentence. Apart from this, whenever we find a morality lexicon word in a training sentence. We replace the word with its synonyms which are present in the same category of the lexicon as the original word. This way we will be able to train the Bert model with multiple lexicon words while keeping the context of the sentence intact. We have used the PyDictionary module to find the synonyms of a moral word.

Each model was trained using Original MFD (MFD) and two different enhanced MFD's (Enh MFD1 and Enh MFD2). This process enabled us to compare different off-the-shelves lexicons and compare their results.

3.4.2. Study 2: Morality and Changing Opinions

For this study, we have taken the individual post and its morality vector and calculated the distance between each of its comments using Euclidean distance by considering the points being on 10-dimensional space. In our observation, we found that a post can have multiple persuasive comments (delta comments) and non-persuasive comments (non-delta comments). To normalize it, we took the mean of all the persuasive comments and non-persuasive comments. While implementing this approach we found that some comments were too short, some posts were too short, comments and posts were deleted, comments and posts did not have a moral context. Due to these reasons, we eliminated the posts which had no persuasive or non-persuasive comment or had a morality vector of 0. The filtering process resulted in a total of 646 original Posts.

To explore the Moral similarity between persuader and persuade, we came up with two Hypotheses:

- Null (H0): The control: The morality of comment have no inference and effect on the Original Poster's opinion
- Alternative (H1): The experimental: Persuasive comment morality is more similar to Original post morality than non-persuasive comment morality.

To test the hypothesis, we used a simulation approach where we used bootstrapping and confidence interval statistical methodologies to check if our hypothesis is consistent with what we observed in the sampling distribution. Sampling distribution helps us create conclusions using statistical results like mean, variance, or standard deviation and Bootstrapping is a method of performing sampling, wherein samples are taken for experimentation, and then put back into the data set to be picked again. For sampling distribution,

we took a sample of 25% of the total data for 100, 000 iteration and calculate the mean in each iteration for the Original Post and persuasive comment distance, Original Post and non-persuasive comment, and the difference between persuasive comment mean and non-persuasive comment mean. In confidence testing on sampling distribution, we took a 95% confidence interval between 2.5 and 97.5 percentile to check if our sampling mean fall between this range.

To validate which hypothesis is correct, we used P-Test. If the distance between Original post (OP) and Persuasive comment is greater than that between OP and Non-Persuasive comment for less than 0.005 times in the sampling distribution we reject the null hypothesis. We applied the test for the Morality Dimension vector, Morality Polarity vector, and Morality type vector using the Bag of word approach.

We also checked how relevant is morality in influencing the Original Poster's opinion by using relevant and temporal grading. For relevance grading, we took the individual posts and sorted the comments based on the euclidean distance from the post in descending order, where the morality vector was calculated using the eMFDScore tool. We took the raw rank of the persuasive comment and calculated the relevance grade using the following formula where N is the total number of comments:

$$RelevanceGrade = \frac{RawRank - 1}{N} * 100 \quad (1)$$

We then plot the frequency distribution of all the Post's Relevance grade. We plot the Temporal grade frequency distribution graph using the same procedure but we sort the comments based on the time of the comment created under each post.

For our next question, are people with stronger moral attitudes more resistant to change? we considered vice and virtue to understand if the original post has stronger moral values, the original poster shows resistance in getting persuaded. Using the Bag of words approach, we calculated the morality vector for each post using the approach stated in study 1. We created heatmaps where the y-axis represents the vice, virtue, and the x-axis represents persuasive comment's location based on its raw rank after all the comments sorted in ascending order based on the time they were created.

3.5. Results

3.5.1. Study 1: Morality and Expressing Opinions

Table 3 and Table 4 show the results of predicting stance when considering morality for traditional machine learning-based model (SVM and RF) and deep learning-based model (LSTM). The Figures also present the accuracy based on original MFD, Enh MFD1, and Enh MFD2.

The results from the Traditional feature-based machine learning models as shown in Table 3 shows that the SVM and RF models were able to predict the stance with the highest accuracy for atheism being 71% and 70% respectively. Using the Enh MFD 1 lexicon, the models were able to predict better compared to Enh MFD 2 and Original MFD in most of the subjects. In Table 4, the LSTM model shows similar results in which Enh MFD 1 dictionary trained model shows the best results for Atheism and Climate subject. A comparison between the two lexicons showed that Enh MFD 1 was able to yield better results for most of the topics in both Traditional and deep learning models. The reason behind this is because Enh MFD 1 was manually annotated and thus making it more reliable for future social computation research.

BERT and XLNet models are able to learn contextual relations between words in a sentence. From Table 5, we can infer that Transformer based models were able to predict the stance when trained using morality focus lexicons better than Traditional feature-based machine learning and deep learning models. When these models were trained with Morality focused sentences, they produced accuracy for feminism at 87% and 89% much higher than the extant literature results [6]. When compared the performance amongst the lexicons, the Enh MFD 1 was able to predict better or on par with other Dictionaries in most of the subjects, which gave us more confidence to use in future research as the results followed the same trajectory as previous models.

3.5.2. Study 2: Morality and Changing Opinions

To validate our first question does morality similarity plays a role in persuasion, we plotted the histogram of the euclidean distance between persuasive comment - Original post (P-O) and non-persuasive comment - Original post (NP-O) represented by red and blue color respectively during each sampling(Figure 2). The red and blue line represents the mean of the Euclidean distance between P-O and NP-O respectively obtained during each sampling of 25% sample size. As the euclidean distance represents the distance between comments and the original post in terms of morality similarity which means the smaller the distance, the closer is the comment and original post in terms of Morality. Inferring from Figure 2, we can say that

Table 3. Results of prediction stance using Traditional feature-based machine learning

Experiments		Stance Dataset											
		Abortion		Atheism		Climate		Clinton		Feminist		Trump	
		SVM	RF	SVM	RF	SVM	RF	SVM	RF	SVM	RF	SVM	RF
Baseline	BaseLine Lexicon	0.61	0.6	0.65	0.7	0.65	0.62	0.53	0.55	0.54	0.56	0.56	0.52
Morality Type	Orig MFD	0.59	0.6	0.71	0.7	0.63	0.63	0.59	0.59	0.53	0.53	0.44	0.41
	Enh MFD 1	0.67	0.62	0.62	0.55	0.47	0.54	0.58	0.61	0.46	0.49	0.37	0.44
	Enh MFD 2	0.57	0.61	0.59	0.58	0.64	0.6	0.64	0.6	0.57	0.57	0.52	0.54
Morality Dimension	Orig MFD	0.52	0.52	0.69	0.69	0.53	0.54	0.6	0.58	0.61	0.61	0.46	0.48
	Enh MFD 1	0.54	0.55	0.57	0.47	0.68	0.63	0.6	0.57	0.56	0.53	0.37	0.35
	Enh MFD 2	0.5	0.48	0.62	0.61	0.6	0.56	0.57	0.57	0.56	0.56	0.44	0.45
Morality Polarity	Orig MFD	0.5	0.5	0.51	0.51	0.61	0.61	0.55	0.55	0.53	0.52	0.31	0.31
	Enh MFD 1	0.7	0.7	0.55	0.55	0.6	0.59	0.6	0.6	0.57	0.57	0.42	0.42
	Enh MFD 2	0.61	0.61	0.68	0.61	0.58	0.6	0.56	0.57	0.51	0.51	0.45	0.48

Table 4. Results of stance prediction using LSTM

Number of Neurons in hidden layer	Approach	Stance Dataset					
		Abortion	Atheism	Climate	Clinton	Feminist	Trump
N = 100	Orig MFD	0.64	0.66	0.69	0.6	0.55	0.5
	Enh MFD 1	0.67	0.71	0.67	0.59	0.53	0.52
	Enh MFD 2	0.63	0.72	0.68	0.57	0.58	0.51
N = 150	Orig MFD	0.69	0.67	0.62	0.58	0.59	0.49
	Enh MFD 1	0.7	0.72	0.71	0.64	0.57	0.51
	Enh MFD 2	0.66	0.71	0.67	0.56	0.56	0.54

the P-O distance is smaller than compared to NP-O which means persuasive comments are more similar to Original post in terms of morality than Non-persuasive comments. Figure 3 represents the histogram where the x-axis values represent the difference between the values of P-O and NP-O during the sampling. In Figure 3, the mean value is -0.21, and the distribution of distance difference is spread over the negative value. This shows that in sampling distribution the majority of the time, the value of P-O was lesser than NP-O.

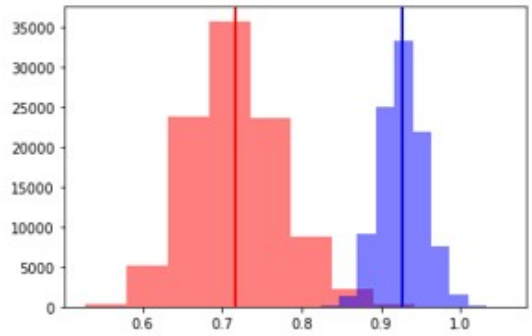

Figure 2. Histogram representation of sampling distribution of P-O and NP-O Euclidean distance

To create the confidence interval of 95%, we cut off the 2.5% lower part (lowest confidence) and 97.5% of the upper part (highest confidence) of the histogram for P-O and NP-O as shown in Figure 3. The main objective of confidence testing is to check if our sampling mean stays between the lowest and highest confidence intervals because that will help in navigating

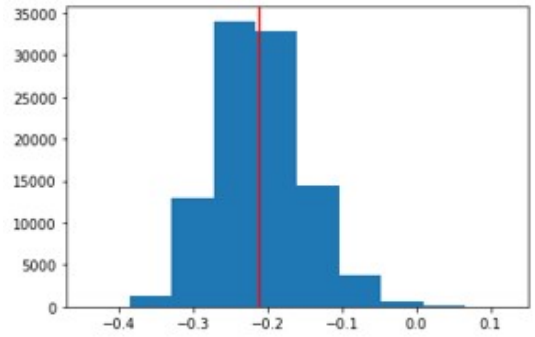

Figure 3. Histogram representing the difference between P-O and NP-O

the uncertainty of how well our sampling distribution has estimated the correct mean value. If the value stays between the lowest and highest confidence interval, we can be confident that the sampling distribution mean is accurate. As shown in Table 6, The sampling mean of P-O is 0.7164 which is between the range of lowest confidence (0.6159) and Highest confidence (0.8391). The same can be observed for NP-O, sampling mean of 0.9268 lies between 0.8751 and 0.9813.

To validate the P-Test, we counted the number of times in a sampling distribution the P-O distance was greater than the NP-O distance and took the mean of it. The value is for the P-test is 0.00124 which is less than 0.005 and hence we can reject the null hypothesis. For the next question, are people with stronger moral attitudes more resistant to change? To observe the resistive nature of strong moral value, we

Table 5. Results of prediction stance using BERT and XLNet

Transformer based models	Moral Dictionary	Abortion	Atheism	Climate	Clinton	Trump	Feminist
BERT	OM	0.79	0.8	0.57	0.8	0.31	0.79
	ENH MFD1	0.67	0.83	0.68	0.83	0.64	0.87
	ENH MFD2	0.66	0.85	0.64	0.84	0.49	0.85
XLNET	OM	0.74	0.84	0.5	0.8	0.38	0.81
	ENH MFD1	0.66	0.77	0.68	0.83	0.75	0.89
	ENH MFD2	0.65	0.82	0.64	0.84	0.61	0.84

Table 6. Confidence interval range and mean value of P-O and NP-O

	Sampling Mean	Lowest Confidence	Highest Confidence
P-O	0.7164	0.6159	0.8391
NP-O	0.9268	0.8751	0.9813

considered Vice Figure 4, Vitrue Figure 5. These heatmaps show the pattern of delta comments based on the morality content of the original post. In Figure 4, if the vice value of the original post is low, the chances of getting delta comments at starting of the conversation are higher. As the moral vice value increases the density of delta comments shifts to the latter of the conversation (with respect to time). The same pattern can be observed in the virtue heatmap. This pattern shows us that if the moral value of an original post is higher than the chances of the comment being persuasive in that starting of conversation decreases which is in line with the previous research.

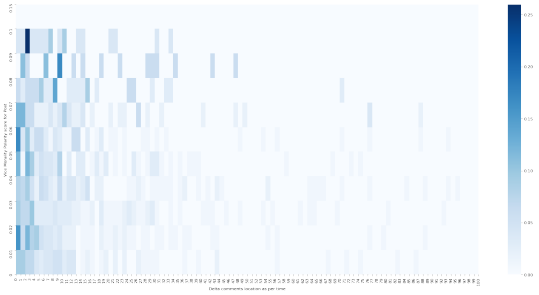

Figure 4. Heatmap representing Original post vice and persuasive comment with the time

4. Discussion

People's moral values are considered to play a key role in their process of forming and/or changing opinions on social and political issues. To gain an empirical understanding of these influences in social media communications, we first examined if actually, the moral value reflect from the text can help us in classifying their opinion better than extant research and secondly, we examined if and how

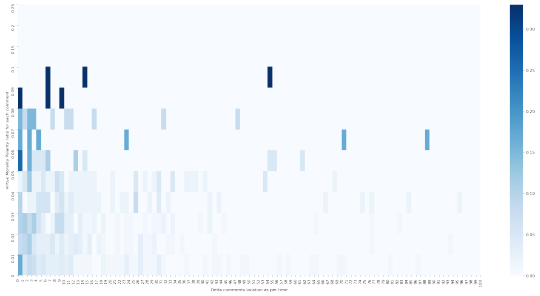

Figure 5. Heatmap representing Original post virtue and persuasive comment with the time

people's morality affects the process of changing their opinions. the morality is related to perceive persuasion or persuasiveness of the online comment. We also observed how strong moral belief is related to once being resistant to online persuasion attempt. These observations have contribute to our understanding of the connection between morality and online reasoning Behavior.

Given our study, Morality from user generated data can help us understand people's moral belief and it helps us to understand people's perspectives, their opinions and the online persuasion processes better. Using these understanding, if the system can detect morality hidden in online data users can be notified about the moral dimensions and the moral prospectus of the other online users that can help them to better understand other people's perspective. Stance detection has been a core component of fact-checking [17], fake news detection [18], and rumor verification [19]. With the help of human value like morality, we can train models better to help us in these tasks.

The limitation of this study is that Enhanced lexicons and tools based on it still have limitation with number of words per category. In future we will be working on a morality lexicon tool using n-gram approach which can help better predict than existing resources. This will pave the path for future morality researchers.

References

[1] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, "Political polarization on twitter," in *Fifth international AAAI*

conference on weblogs and social media, 2011.

[2] A. Öcal, L. Xiao, and J. Park, "Reasoning in social media: insights from reddit "change my view" submissions," *Online Information Review*, 2021.

[3] R. Weber, J. M. Mangus, R. Huskey, F. R. Hopp, O. Amir, R. Swanson, A. Gordon, P. Khooshabeh, L. Hahn, and R. Tamborini, "Extracting latent moral information from text narratives: Relevance, challenges, and solutions," *Communication Methods and Measures*, vol. 12, no. 2-3, pp. 119–139, 2018.

[4] J. Graham and J. Haidt, "The moral foundations dictionary," 2012.

[5] J. A. Frimer, R. Boghrati, J. Haidt, J. Graham, and M. Dehgani, "Moral foundations dictionary for linguistic analyses 2.0," *Unpublished manuscript*, 2019.

[6] R. Rezapour, S. H. Shah, and J. Diesner, "Enhancing the measurement of social effects by capturing morality," in *Proceedings of the tenth workshop on computational approaches to subjectivity, sentiment and social media analysis*, pp. 35–45, 2019.

[7] F. R. Hopp, J. T. Fisher, D. Cornell, R. Huskey, and R. Weber, "The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text," *Behavior Research Methods*, vol. 53, no. 1, pp. 232–246, 2021.

[8] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "Semeval-2016 task 6: Detecting stance in tweets," in *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pp. 31–41, 2016.

[9] L. Tian, X. Zhang, Y. Wang, and H. Liu, "Early detection of rumours on twitter via stance transfer learning," *Advances in Information Retrieval*, vol. 12035, p. 575, 2020.

[10] B. Schiller, J. Daxenberger, and I. Gurevych, "Stance detection benchmark: How robust is your stance detection?," *KI-Künstliche Intelligenz*, pp. 1–13, 2021.

[11] L. Xiao, "A message's persuasive features in wikipedia's article for deletion discussions," in *Proceedings of the 9th International Conference on Social Media and Society*, pp. 345–349, 2018.

[12] J. A. Krosnick and R. E. Petty, "Attitude strength: An overview," *Attitude strength: Antecedents and consequences*, vol. 1, pp. 1–24, 1995.

[13] M. V. Day, S. T. Fiske, E. L. Downing, and T. E. Trail, "Shifting liberal and conservative attitudes using moral foundations theory," *Personality and Social Psychology Bulletin*, vol. 40, no. 12, pp. 1559–1573, 2014.

[14] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto, "Moral foundations theory: The pragmatic validity of moral pluralism," in *Advances in experimental social psychology*, vol. 47, pp. 55–130, Elsevier, 2013.

[15] R. B. Cialdini and R. B. Cialdini, *Influence: The psychology of persuasion*, vol. 55. Collins New York, 2007.

[16] M. Thomas, B. Pang, and L. Lee, "Get out the vote: Determining support or opposition from congressional floor-debate transcripts," *arXiv preprint cs/0607062*, 2006.

[17] J. Thorne, M. Chen, G. Myrianthous, J. Pu, X. Wang, and A. Vlachos, "Fake news stance detection using stacked ensemble of classifiers," in *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pp. 80–83, 2017.

[18] A. Vlachos and S. Riedel, "Fact checking: Task definition and dataset construction," in *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pp. 18–22, 2014.

[19] A. Zubiaga, E. Kochkina, M. Liakata, R. Procter, M. Lukasik, K. Bontcheva, T. Cohn, and I. Augenstein, "Discourse-aware rumour stance classification in social media using sequential classifiers," *Information Processing & Management*, vol. 54, no. 2, pp. 273–290, 2018.

[20] E. Sagi and M. Dehghani, "Measuring moral rhetoric in text," *Social science computer review*, vol. 32, no. 2, pp. 132–144, 2014.

[21] R. Kaur and K. Sasahara, "Quantifying moral foundations from various topics on twitter conversations," in *2016 IEEE International Conference on Big Data (Big Data)*, pp. 2505–2512, IEEE, 2016.

[22] G. M. Fulgoni, A. Lipsman, and C. Davidsen, "The power of political advertising: Lessons for practitioners: How data analytics, social media, and creative strategies shape us presidential election campaigns," *Journal of Advertising Research*, vol. 56, no. 3, pp. 239–244, 2016.

[23] G. Bateson, *Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology*. University of Chicago Press, 2000.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[25] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[26] L. Xiao and T. Khazaei, "Changing others' beliefs online: Online comments' persuasiveness," in *Proceedings of the 10th International Conference on Social Media and Society*, pp. 92–101, 2019.

[27] P. Ben-Nun Bloom and L. C. Levitan, "We're closer than i thought: Social network heterogeneity, morality, and political persuasion," *Political Psychology*, vol. 32, no. 4, pp. 643–665, 2011.

[28] H. Mensah, L. Xiao, and S. Soundarajan, "Characterizing susceptible users on reddit's changemyview," in *Proceedings of the 10th International Conference on Social Media and Society*, pp. 102–107, 2019.

[29] C. McCormick and N. Ryan, "Bert fine-tuning tutorial with pytorch," *Retrieved from*, 2019.

[30] S. Clifford and J. Jerit, "How words do the work of politics: Moral foundations theory and the debate over stem cell research," *The Journal of Politics*, vol. 75, no. 3, pp. 659–671, 2013.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.